Managing Biomedical Science Data: NIH Policy and Practice

Jerry Sheehan
Assistant Director for Policy Development
National Library of Medicine - National Institutes of Health

Workshop on Science Data Management for Government Agencies:
Washington, DC, June 29 – July 1, 2010

Data & information sharing a priority for NIH. . .

Francis S. Collins, M.D., Ph.D.

- Opportunity -- Apply high-throughput technologies to understand fundamental biology and uncover the causes of specific disease states.
- "[High throughput technologies] provide us with the opportunity to ask questions that have the word 'ALL' in them. What are ALL the transcripts in a cell? What are ALL the protein interactions?
- Those kinds of questions are now approachable, especially if we do the right job of ***making really powerful databases publicly accessible to all those who need them*** and empower investigators in small labs as well as big labs to plunge into that kind of mindset."

NIH and Biomedical Data: General Considerations

- Growing volume of biomedical data/information
  - High-throughput genome sequencing
  - High-resolution imaging (e.g., fMRI)
  - Electronic Health Records as source of data for secondary research
- Highly heterogeneous data
  - Genotype, phenotype, imaging, etc. data
  - Often combined in individual studies, e.g., GWAS
- Significant privacy and confidentiality concerns
  - Identifiability of clinical and genomic information
  - Commercial interest
- Wide range of users
  - Scientists, clinicians, patients &families, public health officials
  - Trustworthy information in form tailored to different consumers
- Collected mostly by extramural researchers
  - Researchers define data to be collected and formats
  - Researchers ‘own” the data

NIH Policies for Data/Information Sharing

NIH Public Access Policy

Peer-reviewed manuscript from NIH-funded research in **PubMedCentral** < 12 months after publication

NIH GWAS Policy

Results of NIH-funded GWAS deposited in **dbGaP**

Clinical Trials Reporting

Statutory registration & reporting of summary results of Phase 2-4 drug and device trials: **ClinicalTrials.gov**

NIH Clinical Research Data

Aggregation of **Intramural** data from across ICs into **BTRIS**

IC-specific policies

- NIMH - Autism Research
- NIAAA – Genetics of Alzheimer's
- NIAID – Microbial Genome Sequence
- Others. . .

NIH Data Sharing Policy

Data sharing plan for all application >$500K in any year; Timely release of data

NIH Data Sharing Policy: Basic Elements

- Since October 2003 (FY2004)
- Awards of >$500K in direct costs in a single year
- Requires
  - inclusion of data sharing plan in application – OR --
  - indication of why data cannot be shared (e.g., privacy, national security)
- “Timely release and sharing" = no later than acceptance for publication of the main findings from the final data set
- Plan NOT included in determination of scientific merit or priority score
- Policy does NOT specify where data must be deposited, data formats, etc.

NIH Data Sharing Policy: Recent Developments

- Help grantees better comply with policy
- Centralize information about policy http://sharing.nih.gov
- Develop better guidance for NIH program staff
  - *Suggested Best Practices*
  - http://odoerdb2-1.od.nih.gov/gmac/topics/patents_best_practices.pdf
- Provide additional guidance to grantees
  - Sample data sharing plans
  - Information about NIH-funded repositories
  - Specify *Key Elements* of a data sharing plan (Who, What, Where, When, How) http://grants.nih.gov/grants/sharing_key_elements_data_sharing_plan.pdf.

NIH Genome Wide Association Study (GWAS) Policy

- Effective date – January 25, 2008
- Goals - facilitate broad access to NIH-supported GWAS data to
  - speed translation of basic genetic research into therapies, products, and procedures that benefit the public health.
  - Make more efficient use of GWAS study results – reflecting still high costs of sequencing
- Key elements of policy
  - Applies to all funded studies that include GWAS
  - Applications expected to include a plan for submission of GWAS data to the NIH-designated GWAS data repository (dbGaP)
  - Submit descriptive info including protocol, questionnaires, variables measured – NO STANDARDIZED FORMATS
  - Results as soon as quality control procedures completed – curated, coded phenotype, exposure, genotype data
- Certification that submission consistent w/ policies.

NIH GWAS Policy: Infrastructure -- dbGaP

NIH GWAS Policy: Implementation Considerations

- Privacy
  - Only descriptive data available to public
  - Detailed statistics and patient level data available to qualified researchers
  - Submission must be consistent with informed consent
  - Data Access Committees (DACs) established at participating ICs
  - Data Use Certification by secondary users
- Researcher rights
  - Data available immediately after submission. . . BUT
  - Publication exclusivity for 12 months after deposit of data.
  - Secondary users to acknowledge contributing investigators and funding organizations for original studies
  - ACCESSION NUMBERS at study and variable levels
- Intellectual property rights
  - NIH provides automated calculations of associations
  - Researchers acknowledge goal of ensuring the greatest possible public benefit by signing NIH Data Use Certification

GWAS Policy: Expansion to Other Sequence Data

Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data

Notice Number: NOT-HG-10-006

Key Dates
Release Date: October 19, 2009

Issued by
National Human Genome Research Institute (NHGRI), (http://www.genome.gov/)

Purpose

The purpose of this Notice is to inform the research community of plans by the National Institutes of Health (NIH) to:

1. Update data sharing policies for NIH supported research, including extramural and intramural projects, involving sequence and related genomic data obtained with advanced sequencing technology (*e.g.*, medical resequencing data, sequence data from non-human species, including microorganisms, transcriptomic and epigenomic data, as well as data needed for interpretation, including associated clinical, other phenotype and metadata, such as supporting study documents and methodologies);
2. Encourage investigators and IRBs to consider the potential for broad sharing of sequence and related genomic data in developing informed consent processes and documents for such studies involving human sequence data; and,
3. Communicate the agency's intent and current underlying considerations related to developing a policy pertaining to the deposition of these large datasets into centralized databases, such as the GenBank Short Read Archive (SRA) or the Database of Genotypes and Phenotypes (dbGaP), so that they are available as broadly and rapidly as possible to a wide range of scientific investigators.

Need for Broad Data Sharing Policies

http://grants.nih.gov/grants/guide/notice-files/NOT-HG-10-006.html

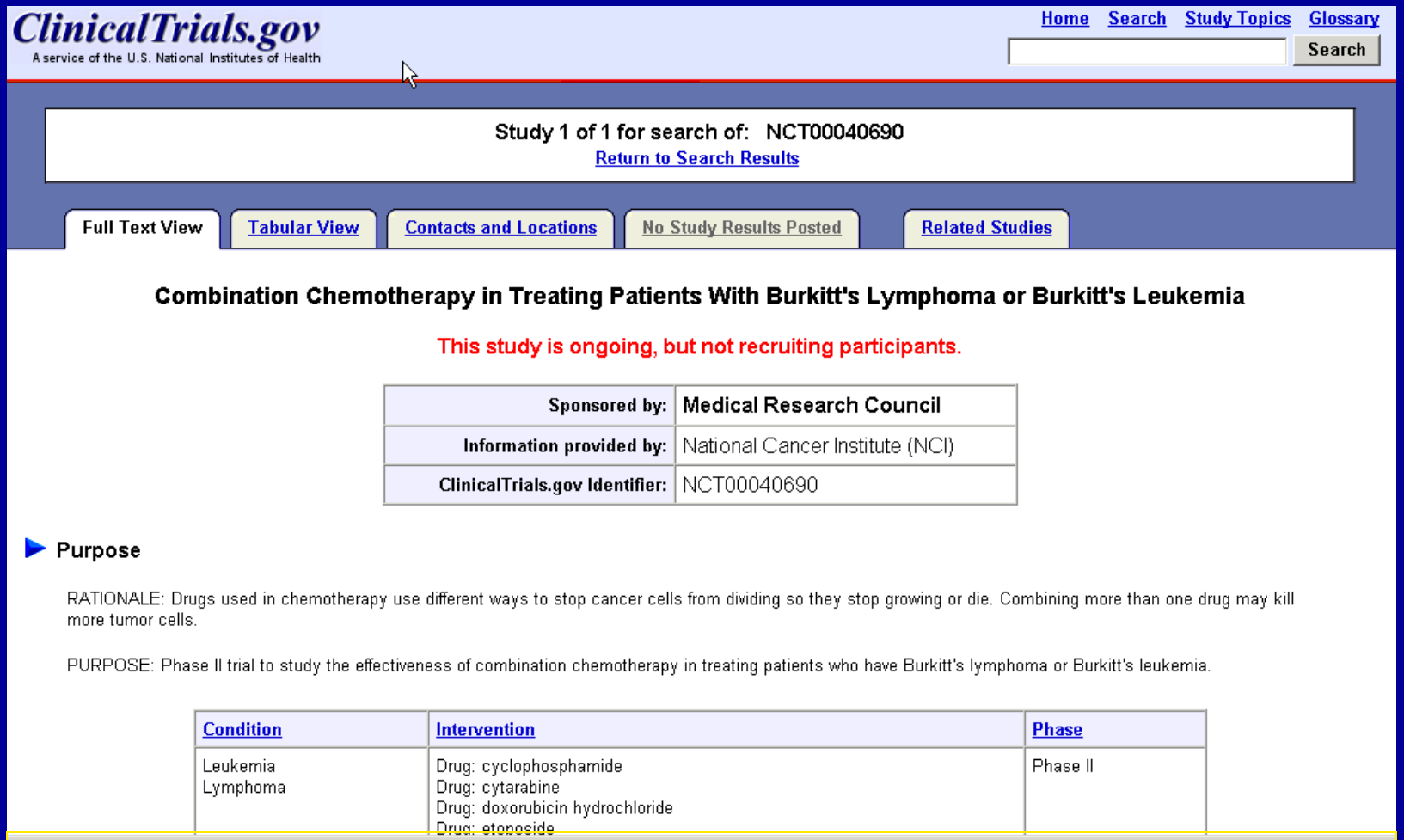
New Infrastructure:
NIH Sequence Read Archive

- Fastest-growing genomic database;
- Currently ~16 terabases + 1 terabase per month (*cf*, Genbank ~100 gigabases)
- Raw sequencing data from next-generation sequencing
- Data exchange with European Read Archive and DDBJ's Read Archive (Japan)

Clinical Trials Registration and Results Information

- Goals
  - Increased transparency of clinical trials (scientific, ethical, care, safety)
  - Address concerns about “hidden results”
  - Facilitate patient enrollment, tracking of results
- Authority
  - FDA Modernization Act of 1997 (FDAMA)
  - FDA Amendments Act of 2007 (FDAAA) – and implementing REGULATIONS
- BASIC REQUIREMENTS
  - Registration of “Applicable Clinical Trials” of drugs and devices within 21 days of first patient enrollment
  - Submit summary results of trials of approved products within 1 year of completion
    - Patient flow and baseline characteristics
    - Primary and secondary outcomes (by arm)
    - Serious and Frequent Adverse Events Expand requirements via Rulemaking
- Enforcement
  - Withholding of grant funds
  - Civil penalties for non-compliance

Clinical Trials Information: Infrastructure – ClinicalTrials.gov



ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Home Search Study Topics Glossary
Search

Study 1 of 1 for search of: NCT00040690
Return to Search Results

Full Text View | Tabular View | Contacts and Locations | No Study Results Posted | Related Studies

Combination Chemotherapy in Treating Patients With Burkitt's Lymphoma or Burkitt's Leukemia

This study is ongoing, but not recruiting participants.

Sponsored by:	**Medical Research Council**
Information provided by:	National Cancer Institute (NCI)
ClinicalTrials.gov Identifier:	NCT00040690

Purpose

RATIONALE: Drugs used in chemotherapy use different ways to stop cancer cells from dividing so they stop growing or die. Combining more than one drug may kill more tumor cells.

PURPOSE: Phase II trial to study the effectiveness of combination chemotherapy in treating patients who have Burkitt's lymphoma or Burkitt's leukemia.

Condition	Intervention	Phase
Leukemia Lymphoma	Drug: cyclophosphamide Drug: cytarabine Drug: doxorubicin hydrochloride Drug: etoposide	Phase II

>90K registered trials (+ 17K per year) and >1800 results

Clinical Trials Information: Implementation Considerations

- Data formats – Tables of Data
  - No established standards for summary level results
  - Must accommodate all trial designs
  - Depositor desribes data and then submits values
- Data curation
  - Data submitted prior to peer review
  - Must be interpreted without narrative text
  - Considerable QA involved and submission of metadata.
- Data attribution
  - UNIQUE IDENTIFIER (NCT #) assigned to each trial.
  - Journal editors require NCT as evidence of registration – precondition for publication
  - CHANGES TRACKED and archived

Clinical Trials Information: Implementation (cont'd)

- Protect patient privacy
  - SUMMARY results information only – not patient-level
- Protect commercial interests:
  - SUMMARY only of protocol and results (to consider submission of full protocol in Rulemaking)
  - Results of APPROVED products only (to consider unapproved products in Rulemaking)
  - Delayed posting of registration information for device trials
  - Specification of data submission requirements

Clinical Trials Information: Next Steps

- Rulemaking
  - Clarify statutory requirements
  - Address “expansion issues” left to Secretary
    - Unapproved products
    - Narrative summaries
    - Other issues. . .
- Education and outreach
  - Grantee community (NIH-funded)
  - Drug and device companies (FDA-regulated)

Biomedical data sharing: Summary observations

- Identify specific data types that are ripe for archiving and sharing
- Link policy development to database development ("If you build it, the will come. . .")
- Use carrots and sticks (mandates accompanied by incentives for sharing, reduced burden)
- Address privacy and confidentiality through tiered levels of access, delayed access, etc.
- Work with relevant communities to develop data standards
- Develop plans for monitoring and enforcement

More information on data & information sharing at NIH

NIH Data Sharing Policies
http://sharing.nih.gov/

NIH GWAS Policy
http://grants.nih.gov/grants/gwas/

ClinicalTrials.gov
http://www.clinicaltrials.gov

NIH Public Access Policy
http://publicaccess.nih.gov/

National Library of Medicine
http://www.nlm.nih.gov